Declarative sequential pattern mining of care pathways*

Thomas Guyet[1], André Happe[2], and Yann Dauxais[3]

[1] AGROCAMPUS-OUEST/IRISA-UMR6074
[2] CHRU Brest/EA-7449 REPERES
[3] Rennes University 1/IRISA-UMR6074

Abstract. Sequential pattern mining algorithms are widely used to explore care pathways database, but they generate a deluge of patterns, mostly redundant or useless. Clinicians need tools to express complex mining queries in order to generate less but more significant patterns. These algorithms are not versatile enough to answer complex clinician queries. This article proposes to apply a declarative pattern mining approach based on Answer Set Programming paradigm. It is exemplified by a pharmaco-epidemiological study investigating the possible association between hospitalization for seizure and antiepileptic drug switch from a french medico-administrative database.

Keywords: answer set programming, epidemiology, medico-administrative databases, patient care pathways

1 Introduction

Pharmaco-epidemiology applies the methodologies developed in general epidemiology to answer questions about the uses of health products in the population in real condition. In pharmaco-epidemiology studies, people who share common characteristics are recruited. Then, a dataset is built from meaningful data (drug exposures, events or outcomes) collected within a defined period of time. Finally, a statistical analysis highlights the links (or the lack of link) between drug exposures and outcomes (*e.g.* adverse effects).

The data collection of such prospective cohort studies is slow and cumbersome. Medico-administrative databases are readily available and cover a large population. They record, with some level of details, all reimbursed drug deliveries and all medical procedures, for insured people. Such database gives an abstract view on longitudinal care pathways. It has become a credible alternative for pharmaco-epidemiological studies [1]. However, it has been conceived for administrative purposes and their use in epidemiology is complex.

Our objective is to propose a versatile pattern mining approach that extracts sequential patterns from care pathways. The flexibility of such new knowledge

* This research is supported by the PEPS project funded by the french agency for health products safety (ANSM) and the SePaDec project funded by Brittany Region.

discovery tools has to enable epidemiologists to easily investigate various types of interesting patterns, *e.g.* frequent, rare, closed or emerging patterns, and possibly new ones. On the other hand, the definition of interesting patterns has to exploit in-depth the semantic richness of care pathways due to complex care event descriptions (*e.g.* units number, strength per unit, drugs and diagnosis taxonomies, etc.). By this mean, we expect to extract less but more significant patterns.

This article presents the application of a declarative pattern mining framework based on Answer Set Programming (ASP) [2] to achieve care pathway analysis answering pharmaco-epidemiological questions.

Answer Set Programming (ASP) is a declarative programming paradigm. It gives a description, in a first-order logic syntax, of what is a problem instead of specifying how to solve it. Semantically, an ASP program induces a collection of so-called *answer sets*. For short, a model assigns a truth value to each propositional atoms of the program. An answer set is a minimal set of true propositional atoms that satisfies all the program rules. ASP problem solving is ensured by efficient solvers. For its computational efficiency, we use *clingo* [3] as a primary tool for designing our encodings. An *ASP program* is a set of rules of the form: $\mathtt{a}_0$ `:-` $\mathtt{a}_1, \ldots, \mathtt{a}_m$, `not` $\mathtt{a}_{m+1}, \ldots,$ `not` $\mathtt{a}_n$, where each $\mathtt{a}_i$ is a propositional atom for $0 \leq i \leq n$ and `not` stands for *default negation*. In the body of the rule, commas denote conjunctions between atoms. If $n = 0$, *i.e.*, the rule body is empty, the rule is called a *fact* and the symbol "`:-`" may be omitted. Such a rule states that the atom a_0 has to be true. If $\mathtt{a}_0$ is omitted, *i.e.*, the rule head is empty, the rule represents an integrity constraint meaning that it must not be *true*. *clingo* also includes several extensions to facilitate the practical use of ASP (variables, conditional literals and cardinality constraints).

Recent researches has been focused on the use of declarative paradigms, including ASP, to mine structured datasets, and more especially sequences [2, 4]. The principle of declarative pattern mining is closely related to the Inductive Logic Programming (ILP) [5] approach. The principle is to use a declarative language to model the analysis task: supervised learning for ILP and pattern mining for our framework. The encoding benefits from the versatility of declarative approaches and offers natural abilities to represent and reason about knowledge.

2 Context, data and pharmaco-epidemiological question

In this work, we exemplify our declarative pattern mining framework by investigating the possible association between hospitalization for seizure and antiepileptic drug switches, *i.e.*, changes between drugs. The first step was to create a digital cohort of 8,379 patients with a stable treatment for epilepsy (stability criterion detailed in [6] have been used). This cohort has been built from the medico-administrative database, called SNIIRAM [1] which is the database of the french health insurance system. It is made of all outpatient reimbursed health expenditures. Our dataset represents 1,8M deliveries of 7,693 different drugs and 20,686 seizure-related hospitalizations.

This dataset and background knowledge (ATC drugs taxonomy, ICD-10 taxonomy) are encoded as ASP facts. For each patient $\boldsymbol{p}$, drug deliveries are encoded with `deliv(`$\boldsymbol{p}$`,`t`,`d`,`q`)` atoms meaning that patient $\boldsymbol{p}$ got q deliveries of drug d at date t. Dates are day numbers starting from the first event date. We use french CIP, Presentation Identifying Code, as drug identifier. The knowledge base links the CIP to the ATC and other related informations (*e.g.* speciality group, strength per unit, number of units or volume, generic/brand-named status, etc). Each diagnosis related to an hospital stay is encoded with `disease(`$\boldsymbol{p}$`,`t`,`d`)` meaning that patient $\boldsymbol{p}$ have been diagnosed with d at date t. Data, $\mathcal{D}$, and related knowledge base, $\mathcal{K}$, represent a total of 2,010,556 facts.

3 Sequential pattern mining with ASP

Let $\mathcal{I} = \{i_1, i_2, \ldots, i_{|\mathcal{I}|}\}$ be a set of *items*. A *temporal sequence* $\boldsymbol{s}$, denoted by $\langle (s_j, t_j) \rangle_{j \in [m]}$ is an ordered list of items $s_j \in \mathcal{I}$ timestamped with $t_j \in \mathbb{N}$. Let $\boldsymbol{p} = \langle p_j \rangle_{1 \leq j \leq n}$, where $p_j \in \mathcal{I}$ be a sequential pattern. We denote by $\mathcal{L} = \mathcal{I}^*$ the *pattern search space*. Given the pattern $\boldsymbol{p}$ and the sequence $\boldsymbol{s}$ with $n \leq m$, we say that $\boldsymbol{s}$ *supports* $\boldsymbol{p}$, iff there exists n integers $e_1 < \ldots < e_n$ such that $p_k = s_{e_k},\ \forall k \in \{1, \ldots, n\}$. $(e_k)_{k \in [n]}$ is called an *embedding* of pattern $\boldsymbol{p}$ in $\boldsymbol{s}$. $\mathcal{E}_{\boldsymbol{p}}^{\boldsymbol{s}} = \{(e_k)_{k \in [n]}\}$ denotes the set of the embeddings of $\boldsymbol{p}$ in $\boldsymbol{s}$. Let $\mathcal{D} = \left\{\boldsymbol{s}^k\right\}_{k \in [N]}$, be a dataset of N sequences. We denote by $\mathcal{T}_{\boldsymbol{p}}$ the sequence set supported by $\boldsymbol{p}$. Given a set of constraints $\mathcal{C}$, the mining of sequential patterns consists in finding out all tuples $\langle \boldsymbol{p}, \mathcal{T}_{\boldsymbol{p}}, \mathcal{E}_{\boldsymbol{p}} \rangle$ satisfying $\mathcal{C}$, where $\mathcal{E}_{\boldsymbol{p}} = \bigcup_{s \in \mathcal{T}_{\boldsymbol{p}}} \mathcal{E}_{\boldsymbol{p}}^{\boldsymbol{s}}$. The most used pattern constraint is the minimal frequency constraint, $c_{f_{min}} : |\mathcal{T}_{\boldsymbol{p}}| \geq f_{min}$, saying that the pattern support has to be above a given threshold f_{min}.

Sequential pattern mining with ASP has been introduced by Guyet et *al.* [2][4]. It encodes the sequential pattern mining task as an ASP program that process sequential data encoded as ASP facts. A sequential pattern mining task is a tuple $\langle \mathcal{S}, \mathfrak{M}, \mathcal{C} \rangle$, where $\mathcal{S}$ is a set of ASP facts encodings the sequence database, $\mathfrak{M}$ is a set of ASP rules which yields pattern tuples from database, $\mathcal{C}$ is a set of constraints (see [4] for constraint taxonomy). We have $\mathcal{S} \cup \mathfrak{M} \cup \mathcal{C} \models \{\langle \boldsymbol{p}, \mathcal{T}_{\boldsymbol{p}}, \mathcal{E}_{\boldsymbol{p}} \rangle\}$.

In our framework, the sequence database is modeled by `seq(`$\boldsymbol{s}$`,`t`,`e`)` atoms. Each of these atoms specifies that the event $e \in \mathcal{I}$ occurred at time t in sequence $\boldsymbol{s}$. On the other hand, each answer set holds atoms that encode a pattern tuples. `pat(`i`,`p_i`)` atoms encode the pattern $\boldsymbol{p} = \langle p_i \rangle_i \in [l]$ where l is given by `patlen(`l`)`, `support(`$\boldsymbol{s}$`)` encodes $\mathcal{T}_{\boldsymbol{s}}$ and finally $\mathcal{E}_{\boldsymbol{p}}$ is encoded by `occ(`$\boldsymbol{s}$`,`i`,`e_i`)` atoms.

4 Declarative care pathway mining

The declarative care pathway mining task can be defined as a tuple of ASP rule sets $\langle \mathcal{D}, \mathfrak{S}, \mathcal{K}, \mathfrak{M}, \mathcal{C} \rangle$ where $\mathcal{D}$ is the raw dataset and $\mathcal{K}$ the knowledge base introduced in section 2; $\mathfrak{M}$ is the encoding of the sequence mining task presented

[4] Original encodings can be found here: https://sites.google.com/site/aspseqmining/

in [2] and $\mathcal{C}$ is a set of constraints. Finally, $\mathfrak{S}$ is a set of rules yielding the sequences database: $\mathfrak{S} \cup \mathcal{D} \cup \mathcal{K} \models \mathcal{S}$. Depending on the study, the expert has to provide $\mathfrak{S}$, a set of rules that specifies which are the events of interest and $\mathcal{C}$, a set of constraints that specifies the patterns the user would like.

In the following of this section, we give examples for $\mathfrak{S}$ and $\mathcal{C}$ to design a new mining tasks inspired from a *case-crossover study* answering our clinical question [6]. For each patient, the $\mathfrak{S}$ rules generate two sequences made of deliveries within respectively the 3 months before the first seizure (positive sequence) and the 3 to 6 months before the first seizure (negative sequence). In this setting the patient serves as its own control. The mining query consists in extracting frequent sequential patterns where a patient is supported by the pattern iff the pattern appears in its positive sequence, but not in its negative sequence. A frequency threshold for this pattern is set up to 20 and we also constraint patterns 1) to have generic and brand-name deliveries and 2) to have exactly one switch from a generic to a brand-name anti-epileptic grugs – AED (or the reverse).

Defining sequences to mine with $\mathfrak{S}$. Listing above illustrates the sequence generation of deliveries of anti-epileptic drug specialities within the 3 months (90 days) before the first seizure event. It illustrates the use of the knowledge base to express complex sequences generation. In this listing, `aed(`i`,`c`)` lists the CIP code i, which are related to one of the ATC codes for AED ($N03AX09$, $N03AX14$, etc.), and `firstseizure(`$\boldsymbol{p}$`,`t`)` is the date, t, of the first seizure of patient $\boldsymbol{p}$. A seizure event is a disease event with one of the G40-G41 ICD-10 code. The first seizure is the one without any other seizure event before. ASP enables to use a reified model of sequence where events are functional literals. `seq(P,T,deliv(AED,Gr,G))` designates that patient `P` was delivered at time `T` with a drug where `AED` is the ATC code, `Gr` identify the drug speciality and `G` indicates whether the speciality is a generic drug or a brand-named one. The same encoding can be adapted for sequences within the 3 to 6 months before the first seizure event.

```
aed(CIP,AED):-cip_atc(CIP,AED),AED=(n03ax09;n03ax14;n03ax11;n03ag01;n03af01).
firstseizure(P,T) :- disease(P,T,D), is_a(D,g40;g41),
                     #count{Tp: disease(P,Tp,Dp), is_a(Dp,g40;g41), Tp<T}=0.

seq(P,T,deliv(AED,Gr,1)):- deliv(P,T,CIP,Q), aed(CIP,AED), grs(CIP,Gr),
                           generic(CIP), T<Ts, T>Ts-90, firstseizure(P,Ts).

seq(P,T,deliv(AED,Gr,0)):- deliv(P,T,CIP,Q), aed(CIP,AED), grs(CIP,Gr),
                        not generic(CIP), T<Ts, T>Ts-90, firstseizure(P,Ts).
```

Defining constraints on patterns. On the other side of our framework, $\mathcal{C}$ enables to add constraints on patterns the clinician looks for. Lines 1-3 (see listing above) encode the case-crossover constraints. They select patterns (*i.e.*, answer sets) that are frequently in the 3 months period but not in the 3 to 6 months period. The frequency threshold is set to 20. Finally, lines 5-6 illustrate a constraint on the shape of the pattern, that here must contains exactly one switch from a brand-name to a generic drug (or the reverse).

```
1 discr(T) :- support(T), not neg_support(T).
2 #const th=20.
```

```
3 :- { discr(T) } < th.
4
5 change(X) :- pat(X+1,deliv(AEDp,GRSp,Gp)), pat(X,deliv(AED,GRS,G)), Gp!=G.
6 :- #count{X:change(X)}!=1.
```

Results The solver extracts respectively 32 patterns and 21 patterns (against 4359 patterns with a regular sequential pattern mining algorithm). With such very constrained problem, the solver is very efficient and extracts all patterns in less than 30 seconds. The following pattern is representative of our results: $\langle(N03AG01, 438, 1), (N03AG01, 438, 1), (N03AX14, 1023, 0), (N03AX14, 1023, 0)\rangle$ is a sequence of deliveries showing a change of treatment from a generic drug of the speciality 438 of valproic acid to the brand-name speciality 1023 of levetiracetam. According to our mining query, we found more than 20 patients which have this care sequence within the 3 months before a seizure, but not in the 3 previous months preceding this period. These new hypothesis of care-sequences are good candidates for further investigations and possible recommendation about AE treatments.

5 Conclusion

Declarative sequential pattern mining with ASP is an interesting framework to flexibly design care-pathway mining queries that supports knowledge reasoning (taxonomy and temporal reasoning). We illustrated the expressive power of this framework by designing a new mining tasks inspired from case-crossover studies and shown its utility for care pathway analytics. We strongly believe that our integrated and flexible framework empowers the clinician to quickly evaluate various pattern constraints and that it limits tedious pre-processing phases.

References

1. Martin-Latry, K., Bégaud, B.: Pharmacoepidemiological research using french reimbursement databases: yes we can! Pharmacoepidemiology and drug safety **19**(3) (2010) 256–265
2. Gebser, M., Guyet, T., Quiniou, R., Romero, J., Schaub, T.: Knowledge-based sequence mining with ASP. In: Proceedings of IJCAI. (2016) 1497–1504
3. Gebser, M., Kaminski, R., Kaufmann, B., Ostrowski, M., Schaub, T., Schneider, M.: Potassco: The Potsdam answer set solving collection. AI Communications **24**(2) (2011) 107–124
4. Negrevergne, B., Guns, T.: Constraint-based sequence mining using constraint programming. In: Proceedings of International Conference on Integration of AI and OR Techniques in Constraint Programming, CPAIOR. (2015) 288–305
5. Quiniou, R., Cordier, M.O., Carrault, G., Wang, F.: Application of ILP to cardiac arrhythmia characterization for chronicle recognition. In: Proceedings of conference on Indictive Logic Programming. (2001) 220–227
6. Polard, E., Nowak, E., Happe, A., Biraben, A., Oger, E.: Brand name to generic substitution of antiepileptic drugs does not lead to seizure-related hospitalization: a population-based case-crossover study. Pharmacoepidemiology and drug safety **24**(11) (2015) 1161–1169